Interrater and intrarater reliability of the Knosp scale for pituitary adenoma grading

Michael A. Mooney, MD, Douglas A. Hardesty, MD, John P. Sheehy, MD, Robert Bird, MD, Kristina Chapple, PhD, William L. White, MD, and Andrew S. Little, MD

Department of Neurosurgery, Barrow Neurological Institute, St. Joseph’s Hospital and Medical Center, Phoenix, Arizona

OBJECTIVE The goal of this study was to determine the interrater and intrarater reliability of the Knosp grading scale for predicting pituitary adenoma cavernous sinus (CS) involvement.

METHODS Six independent raters (3 neurosurgery residents, 2 pituitary surgeons, and 1 neuroradiologist) participated in the study. Each rater scored 50 unique pituitary MRI scans (with contrast) of biopsy-proven pituitary adenoma. Reliabilities for the full scale were determined 3 ways: 1) using all 50 scans, 2) using scans with midrange scores versus end scores, and 3) using a dichotomized scale that reflects common clinical practice. The performance of resident raters was compared with that of faculty raters to assess the influence of training level on reliability.

RESULTS Overall, the interrater reliability of the Knosp scale was “strong” (0.73, 95% CI 0.56–0.84). However, the percent agreement for all 6 reviewers was only 10% (26% for faculty members, 30% for residents). The reliability of the middle scores (i.e., average rated Knosp Grades 1 and 2) was “very weak” (0.18, 95% CI 0.27 to 0.56) and the percent agreement for all reviewers was only 5%. When the scale was dichotomized into tumors unlikely to have intraoperative CS involvement (Grades 0, 1, and 2) and those likely to have CS involvement (Grades 3 and 4), the reliability was “strong” (0.60, 95% CI 0.39–0.75) and the percent agreement for all raters improved to 60%. There was no significant difference in reliability between residents and faculty (Residents 0.72, 95% CI 0.55–0.83 vs faculty 0.73, 95% CI 0.56–0.84). Intrarater reliability was moderate to strong and increased with the level of experience.

CONCLUSIONS Although these findings suggest that the Knosp grading scale has acceptable interrater reliability overall, it raises important questions about the “very weak” reliability of the scale’s middle grades. By dichotomizing the scale into clinically useful groups, the authors were able to address the poor reliability and percent agreement of the intermediate grades and to isolate the most important grades for use in surgical decision making (Grades 3 and 4). Authors of future pituitary surgery studies should consider reporting Knosp grades as dichotomized results rather than as the full scale to optimize the reliability of the scale.

https://thejns.org/doi/abs/10.3171/2016.3.JNS153044

KEY WORDS cavernous sinus; interrater reliability; intrarater reliability; Knosp classification; pituitary adenoma; pituitary surgery
been shown to predict the intraoperative finding of CS invasion and have been inversely correlated with gross-total resection and endocrinological remission rates.\textsuperscript{25}

Although the Knosp grading scale has been widely applied to pituitary research studies, its reliability and agreement among raters have not been investigated. In this study, we investigated the interrater and intrarater reliability of the Knosp grading scale to improve future reporting of pituitary research studies. We also evaluated a commonly reported dichotomization of the scale, which divides tumors into those likely to have CS invasion (Grades 3 and 4) and those unlikely to have CS invasion (Grades 0–2). Lastly, we assess the effect of training level on the reliability of the scale.

**Methods**

**Patients and Raters**

Fifty unique, gadolinium-enhanced, dedicated pituitary MRI studies were selected from a prospectively maintained database of biopsy-proven, non-functioning pituitary macroadenomas. All studies were preoperative MRIs obtained in patients who had not undergone prior surgery. Scans were preliminarily evaluated by a single author (M.A.M.) to ensure that a range of tumor sizes and degrees of CS involvement were included. Six independent raters participated in the study: 3 neurosurgery residents (M.A.M., D.A.H., and J.P.S.) and 3 faculty raters (2 pituitary surgeons [W.L.W. and A.S.L.] and 1 neuroradiologist [R.B.]). Each rater was given a written and verbal description of the project and the original manuscript describing the Knosp grading scale for reference. Three raters participated in the intrarater reliability portion of the study (D.A.H., J.P.S., and A.S.L.) and were presented with the MRI scans the second time more than 4 weeks after their initial review. For the intrarater reliability portion of the study, scans were de-identified to avoid rater recognition. This study was approved by the institutional review board of St. Joseph’s Hospital and Medical Center, Phoenix, Arizona.

**Statistics**

The number of scans analyzed and the number of participants included were determined based upon a pre-study power analysis. Briefly, a sample of 25 scans was necessary to detect a significant correlation coefficient of 0.60 with p < 0.05 and 90% power. This number was doubled to 50 scans to allow for subgroup analyses. The intrarater analysis was performed with 3 physicians on 35 scans and had 80% power to detect a difference between 2 correlations based on a large effect size of Cohen’s q equal to 0.638.

Pairwise Spearman correlation coefficients were calculated among raters and averaged to describe the scale reliability. The 6 ratings for each MRI scan were then averaged and sorted in ascending order. Average rated scores less than 1 or greater than 2 were considered the “end” scores, and average rated scores between 1 and 2 were considered the “middle scores.” Spearman correlation coefficients using the middle and ends of the scale were then calculated to determine whether physician reliability varied at different points on the scale. The Spearman correlation coefficient was chosen over Cohen’s kappa because the Knosp grading scale is ordinal and not strictly categorical. Although not shown in this paper, kappa coefficients were also calculated and, as expected, were generally lower than the Spearman coefficients.

We also tested a commonly reported dichotomization (Grades 0–2 vs Grades 3 and 4). The average of the pairwise phi coefficients for all 6 raters was used as an overall reliability for the dichotomous scale. Dichotomous scale reliabilities for resident and faculty physicians were summarized using the phi coefficient for 2 raters. The phi coefficient was chosen because it is a measure of association between binary variables and is a correlation coefficient in the 2 × 2 case, thus similar to the Spearman correlation coefficient in our study. Counts and percentages representing rater agreement are also presented. Coefficients were interpreted as follows: 0.00–0.19 indicates “very weak”; 0.20–0.39, “weak”; 0.40–0.59, “moderate”; 0.60–0.79, “strong”; and 0.80–1.00, “very strong.” All calculations were performed by a dedicated biostatistician (K.C.) using SPSS Statistics for Windows, Version 22.0 (IBM Corp.).

**Results**

**Interrater Reliability**

Overall interrater reliability for the full Knosp grading scale was “strong” (0.73, 95% CI 0.56–0.84) (Table 1). However, when examined separately, reliability for the middle scores (i.e., average rated score between 1 and 2) was “very weak” (0.18, 95% CI 0.27 to 0.56). The reliability of the scale ends (i.e., average rated score of less than 1 or greater than 2) was “very strong” (0.84, 95% CI 0.71–0.93). Table 1 presents the interrater reliability for the Knosp grading scale.

**Table 1. Interrater reliability of the Knosp scale**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Reliability (95% CI)</th>
<th>Percent Agreement (no./total [%])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability of the full scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All raters &amp; scans (6 raters)</td>
<td>0.73 (0.56–0.84)</td>
<td>5/50 (10%)</td>
</tr>
<tr>
<td>Faculty raters (3 raters)</td>
<td>0.73 (0.56–0.84)</td>
<td>13/50 (26%)</td>
</tr>
<tr>
<td>Resident raters (3 raters)</td>
<td>0.72 (0.55–0.83)</td>
<td>15/50 (30%)</td>
</tr>
<tr>
<td>Reliability of middle scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All raters (6 raters)</td>
<td>0.18 (−0.27 to 0.56)</td>
<td>1/22 (5%)</td>
</tr>
<tr>
<td>Faculty raters (3 raters)</td>
<td>0.22 (−0.22 to 0.59)</td>
<td>6/22 (27%)</td>
</tr>
<tr>
<td>Resident raters (3 raters)</td>
<td>0.15 (−0.29 to 0.54)</td>
<td>5/22 (23%)</td>
</tr>
<tr>
<td>Reliability of scale ends</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All raters (6 raters)</td>
<td>0.84 (0.69–0.93)</td>
<td>4/28 (14%)</td>
</tr>
<tr>
<td>Faculty raters (3 raters)</td>
<td>0.85 (0.69–0.93)</td>
<td>7/28 (25%)</td>
</tr>
<tr>
<td>Resident raters (3 raters)</td>
<td>0.86 (0.71–0.93)</td>
<td>10/28 (36%)</td>
</tr>
<tr>
<td>Dichotomous Knosp scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All raters &amp; scans (6 raters)</td>
<td>0.60 (0.39–0.75)</td>
<td>30/50 (60%)</td>
</tr>
<tr>
<td>Faculty raters (3 raters)</td>
<td>0.60 (0.39–0.75)</td>
<td>38/50 (76%)</td>
</tr>
<tr>
<td>Resident raters (3 raters)</td>
<td>0.56 (0.34–0.73)</td>
<td>34/50 (68%)</td>
</tr>
</tbody>
</table>
CI 0.69–0.93) (Fig. 1). When the scale was dichotomized into clinically useful groups of grades likely to indicate CS invasion (Grades 3 and 4) versus those unlikely to indicate CS invasion (Grades 0–2), the scale reliability was “moderate” to “strong.” The percent agreement greatly improved for the dichotomized scale compared with the full scale, as shown in Fig. 2 and Table 1 (increasing from 26% to 76% for faculty, from 30% to 68% for neurosurgery residents, and from 10% to 60% for residents and faculty combined). Of note, the overall decrease in reliability with dichotomization is expected, given the smaller scale range included in the statistical analysis.6,28

Intrarater Reliability

Three raters (1 faculty member and 2 residents) reviewed and rated 35 of the scans twice to determine intrarater reliability. Spearman correlation coefficients between the first and second ratings for the 2 resident surgeons were in the “moderate range” (0.55, 95% CI 0.26–0.74 and 0.53, 95% CI 0.25–0.74), whereas the correlation for the faculty surgeon was strong (0.79, 95% CI 0.62–0.89). The percent agreement for the 0–4 scale for both measurements was 42.9% (n = 15) and 65.7% (n = 23) for resident raters and 65.7% (n = 23) for the surgeon rater.

In the dichotomized form of the scale, the percent agreement increased to 77.1% (n = 27) for both resident raters and 91.4% for the faculty rater.

Training Level

There were no statistically significant differences in scale interrater reliability with respect to training level (Table 1, Fig. 1). Faculty and resident raters had similar reliability for the full scale when all MRI scans were included (0.73 vs 0.72, respectively; p = 0.92); the reliability remained similar when only the middle scores were considered (0.22 vs 0.15, p = 0.80) and when only the end scores were considered (0.85 vs 0.86, p = 0.91). For the dichotomous scale, there was also no statistically significant difference between faculty and resident reliabilities (0.60 vs 0.56, p = 0.77).

Intrarater reliability was higher for the faculty rater than for either of the resident raters, although the difference was not statistically significant (p = 0.064 for Rater 1 and p = 0.056 for Rater 2) (Table 2). In the dichotomized form of the scale, the intrarater reliability for the faculty rater was significantly higher than for resident Rater 1 (p = 0.003) and resident Rater 2 (p = 0.046).

Discussion

Assessing the reliability of a proposed grading scale is essential for validation and implementation of the scale. Numerous scales used in neurosurgery have been evaluated over the past decade in attempts to validate their utilization for both clinical decision making and research.2,7,13,15–17,19,23,33,34 Key features include accuracy and reproducibility, which can be estimated by studying interrater and intrarater reliability.

Preoperative imaging of pituitary adenomas can be used to predict the invasiveness of a tumor and the likelihood of achieving a surgical cure, which impacts pre-
operative patient counseling. As pituitary adenomas enlarge, they can erode the sellar floor and expand into the sphenoid sinus, suprasellar space, and CS. The presence of CS invasion as estimated on preoperative imaging is predictive of incomplete resection and decreased cure rates.

Since the early efforts utilizing CT to evaluate parasellar extension and CS invasion, gadolinium-enhanced MRI has replaced CT as the superior imaging study for assessing the CS space. The addition of gadolinium enhances the venous plexus within the CS, which allows for examination of the relation of the tumor to the medial CS wall. In 1993, Knosp et al. proposed a grading scale for assessing the likelihood of CS invasion based on an adenoma’s relation to the cavernous ICA and CS venous plexus, as seen on gadolinium-enhanced MRI. In this system, the adenoma’s parasellar extension in relation to the cavernous ICA determines the grade (0–4), with higher grades demonstrating an increased likelihood of CS invasion. Interestingly, in that original publication, 88% of Grade 2 and 100% of Grade 3 lesions were confirmed to have invaded the CS at the time of surgery. Since that time, the implementation of the endoscope has allowed for improved visualization of the CS medial wall in some cases, and the surgically confirmed rates of invasion are significantly lower than initially reported (9.9% of Grade 2 lesions; 37.9% of Grade 3). In addition to its intended use for predicting the likelihood of CS invasion, the Knosp grading scale has been widely implemented for classifying pituitary adenomas for various research studies.

Despite the widespread use of Knosp grading as a standard scale, its interrater and intrarater reliability has not been evaluated. Our analysis has several notable findings. First, it demonstrated that the interrater reliability of the middle scores (i.e., average Knosp Grades 1 and 2) of the scale was “very weak.” In fact, the percent agreement for all raters was only 5% for these middle scores. This has significant implications for interpreting past research studies and guiding future studies of pituitary adenomas, as tumor grades in the middle of the Knosp scale are not reliable. Published studies utilizing the full Knosp scale must therefore be interpreted with caution.

The second notable finding is that we tested a commonly reported dichotomization of the scale into grades unlikely (Grades 0–2) and likely (Grades 3 and 4) to represent CS invasion. When we dichotomized Knosp grading into these two clinically useful groups, we found the interrater reliability to be “moderate” to “strong.” Furthermore, the percent agreement between raters improved from 10% for the full scale to 60% for the dichotomized scale. These findings support studies that have used this type of dichotomization, and we encourage future authors to employ this strategy.

The third notable finding of this study is that training level did not have a significant effect on interrater reliability. This finding is of interest because trainees are often involved in pituitary research studies. As long as residents receive appropriate training on how to use the scale, their scoring of Knosp grade can be considered reliable, although intrarater reliability may be improved with more experienced raters.

The main limitation of this study, and any study of this type, is in the selection of raters. The grading of MRI scans by the pituitary surgeons, neurosurgery residents, and neuroradiologists is assumed to be generalizable to other raters with similar skill levels. It should be noted that this study was performed at a high-volume pituitary center and that it included more senior neurosurgery residents (postgraduate years 3–5). A neuroradiologist and 2 pituitary surgeons were included to provide the highest possible standard for preoperative Knosp grading. All raters were provided with a description of the study and the original manuscript describing the grading scale. We purposefully did not have training sessions for the raters. In this way, we sought to recapitulate the “real world” application of the grading scale.

The second limitation is that we did not analyze the most recently proposed grading scale by Micko et al. This scale subdivides Knosp Grade 3 into Grade 3A and Grade 3B on the basis of superior or inferior extension within the CS, and substantial differences were found in CS invasion rates. We chose not to evaluate this modification of the Knosp scale in this study because it is not as commonly reported as the original Knosp scale. We anticipate that the reliability and percent agreement for the modified scale would be lower, given its increased complexity. Future studies designed to evaluate the modified scale are warranted.

The third limitation of this study is that we did not compare preoperative Knosp grades with intraoperative findings, which has been done elsewhere. Our study examined the reliability of preoperative Knosp grading for categorizing pituitary adenomas, not the positive/negative predictive value of the Knosp scale. The Knosp scale is commonly used as a descriptor of pituitary adenomas in research studies and surgical series, and our goal was to determine the validity of this practice. The limitations of the Knosp scale for accurately predicting CS invasion and alternative proposals for classification based on intraoperative findings have been detailed elsewhere and should be considered by future authors of pituitary studies.

**Conclusions**

While our findings suggest that the Knosp grading scale has acceptable interrater reliability overall, this study...
raises important questions about the “very weak” reliability of the scale’s intermediate grades and the poor percent agreement for the full scale. By dichotomizing the scale into clinically useful groups, we were able to address the poor reliability of the intermediate grades and isolate the grades that are most important for surgical decision making (Grades 3 and 4). Authors of future pituitary surgery studies should consider reporting Knosp grades as dichotomized results rather than using the full scale to improve agreement among raters across different studies.

References


Disclosures
The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author Contributions
Conception and design: Little, Mooney, Chapple. Acquisition of data: Little, Mooney. Analysis and interpretation of data: all authors. Drafting the article: Little, Mooney, Chapple. Critically revising the article: Little, Mooney, Chapple. Reviewed submitted version of manuscript: Little, Mooney, Chapple. Approved the final version of the manuscript on behalf of all authors: Little. Statistical analysis: Little, Chapple. Administrative/technical/material support: Little, White. Study supervision: Little.

Correspondence
Andrew S. Little, c/o Neuroscience Publications, Barrow Neurological Institute, St. Joseph’s Hospital and Medical Center, 350 W Thomas Rd., Phoenix, AZ 85013. email: neuropub@dignityhealth.org.